ROMANIAN ASSOCIATIVE DICTIONARY

VICTORIA BOBICEV, VICTORIA MAXIM

Technical University of Moldova, Computers, Informatics and Microelectronics Faculty

victoria_bobicev@rol.md *maxivica@yahoo.com*

Abstract

The paper reports about an experiment of creation and development of an associative dictionary for the Romanian language. It outlines the first phase of the experiment when word associations were collected using questionnaire surveys. The second phase includes online interface creation and expanding the dictionary via internet. Several technical issues of the second phase are discussed. Some preliminary statistics are presented and a brief analysis of the obtained database is made. The created dictionary can be used in lexicography and studying Romanian language. At this stage of work we however are more interested in the richest and the most representative database of word association; the detailed analysis is postponed to forthcoming research.

1. Introduction

Semantics is proven to be the nuclear element of language. Lexical semantic networks are of great importance in Computational Linguistics of our days. The WordNet (Miller, 1995) wide popularity is the argument which proves the utility of semantic lexicons. One of the WordNet shortcomings is a small number of semantic relations. Other semantic lexicons like EuroWordNet[1] and Simple[2] were created to solve this problem. Semantic relations in these lexicons are well considered by the competent linguists and based on various lexical theories.

Our lexicon is created relying on different principles. The source of relations is the first main difference. Relations between words are obtained directly from the native speakers of the language as their free associations. The second main difference is the type of the involved relations. We do not name these relations or classify them; these are just relations of free associations in the human's mind. In psychology, free associations are the first words that come to the mind of a native speaker when he or she is presented with a stimulus word, presumably retrieved from associative memory (Nelson et al., 1998). The word presented to respondent is called "stimulus" and the word that comes to the mind is called "response". This type of relations is being studied in various domains of research, such as psychology, artificial intelligence, computational linguistics and natural language processing.

Associative dictionary is a collection of the word pairs "stimulus - response" and represents the language in a somehow unusual form - not in the form of continuous text, as in a novel or newspaper article, not in the form of a systematic description, as in

[1] http://www.illc.uva.nl/EuroWordNet/
[2] http://www.ub.edu/gilcub/SIMPLE/reports/simple/Site_simple.htm

grammar or dictionary, but as pair (combination) of words or word groups that serve as building material for the detailed phrases for constructing sentences.

Thus, any word in our minds, in memory, just like in the speech chain, does not exist in isolation. Any word requires a kind of "extension", looking for its pair, wants to become "a model of two words." And such possible "extensions", such models of two words - typical, easily reproducible, believable and understandable to native speaker - are recorded in the associative dictionary. In addition, each pair of stimulus-response is not a complete sentence, but a necessary component of it – it is either a grammatically formalized part, or only the core of a future statement which will be given the completed form (Уфимцева 2004).

The paper reports on an experiment of the word's associations for Romanian database creation. It outlines the first phase of the experiment while word associations were collected using questionnaire surveys. Next, the second phase is described which include online interface creation and augmentation of the dictionary via internet. Several technical issues of the second phase are discussed. Some preliminary statistics is presented and a brief analysis of the obtained database is made. At this stage of the work, we are interested in the richest and the most representative database of word association; the detailed analysis is postponed for the upcoming research.

2. Related work

There are a number of semantic lexicons with a various relations between words. The most popular is WordNet which contain a relatively small number of relations; it is considered one of its disadvantages. EuroWordNet authors revised and enlarged this set of relations. Simple uses Qualia structure theory as a source of semantic relations in lexicon (Pustejovsky, 2010). The attempt to code as much relations as possible has its negative effect; these lexicons are difficult to process. Fairly sophisticated algorithms are required to obtain the necessary information in a plausible time.

Knowledge bases are the other types of semantic information sources. Well-known CYC[3] include the lexicon as part of the knowledge base. Words in the lexicon are connected with knowledge base concepts thus obtaining semantic capacity. The number of concepts and relations is one of the largest among various resources of this kind. On the contrary, ConceptNet[4] describes only 20 types of relations; some of them are similar with other resources. It is the only resource which is created not by skilled linguists but by volunteers via online interface. This method of knowledge acquisition has several advantages: no need of professional linguists with special training, which leads to less cost and higher growing rate.

Associations between words are obtained also from people without any Special knowledge of linguistics; the only demand is that they should be native speakers of language. Though word association experiments are a usual psychological practice, the obtained results are of great interest in various domains of research, as for example in cognitive science. The most important among these is the understanding that the association is a fundamental mechanism underlying human knowledge (Cramer, 1968,

[3] http://www.cyc.com/
[4] http://conceptnet5.media.mit.edu/

Dees, 1965). This notion is compatible with a number of statements in the field of natural language processing research such as the notion of mutual information (Church and Hank, 1990) as a measure of the importance of an association between two words (Hirst, 2004) and confirmation of the fact that a lexicon can often be a useful basis for the creation of practical ontologies. Lexical networks, represented by lexical nodes (Collins, Loftus, 1975) are the basic points of many connection patterns of human thoughts.

Recently, word associations have being studied by a number of researchers in the domain of cognitive science (Nelson et al., 2005; Steyvers et al., 2004). All these studies use The University of South Florida *Word Association, Rhyme, and word fragment Norms* (Nelson et al., 1998), which is the largest database of American English words associations, comprising nearly 5,000 words and a response average of 149 for each word collected from more than 6000 participants during the years 1975-2000.

There are various sources of word associations for different languages. The already mentioned largest database of word association for English[5]. We should also mention Edinburgh Associative Thesaurus (Kiss et al, 1973) freely available database for English[6]. Among the resources for other languages the Russian associative dictionary (Караулов, 2003), Bulgarian associative dictionary (Балтова 2003), the integrated Slavic dictionary (Уфимцева, 2004) are to be referred. All these resource were collected manually using questionnaire surveys. The more recent resources have been creating using online interface are the Large-Scale Database of Japanese Word Associations (Joyce, 2005), French associative dictionary[7], word association game for English[8], online interface for Russian associative dictionary[9].

3. The first step of Romanian word associations database creation

The first collection of Romanian word associations was created by the direct interrogation. 150 stimulating words were selected from the list of the most frequent Romanian words. The frequency list was created for the corpus described in (Vlad, 2005). The corpus was created on the base of 93 books of various genres: Romanian and foreign fiction, religious literature, philosophy, medical texts, history, law, and others. The authors' aim was to include in the corpus as much types of literature as it was possible. The corpus overall volume is 8.8 million words; the corpus frequency dictionary consists of more than 200 000 words. It is well known that the most frequent words in text are so called "stop-words": articles, prepositions, conjunctions, pronouns and some others which do not carry much semantic information and are used for syntactically correct sentences formation. We obviously were not interested in these words; we selected the most frequent 50 nouns, 50 adjectives and 50 verbs. This list of 150 words arranged in the first column of a table was presented to respondents. They

[5] freely available at http://w3.usf.edu/FreeAssociation
[6] http://www.eat.rl.ac.uk/
[7] http://dictaverf.nsu.ru/fr
[8] http://wordassociation.org
[9] http://tesaurus.ru/dict/dict.php

had to write in the second column of the table the word they were associating in mind while reading the word from the first column of the table.

The respondents were 50 students aged between 19-21 years. Each of them was given a MSWord document with the described above table and they completed the second column of the table. We were interested in the statistical results and the inquiries were anonymous.

Table 1: The strongest associations from the Romanian word association database.

Stimulating word	Association	Number of respondents	Number of respondents providing this association
forța	putere	50	29
ciudat	straniu	50	22
ceas	timp	50	21
noaptea	întuneric	50	21
bucurie	fericire	50	18
istoria	trecut	50	18
târziu	noapte	50	18
moment	clipa	50	17
nevoie	necesitate	50	17
bucătărie	mîncare	50	15
frig	iarna	50	15
piatra	tare	50	15

The obtained data was analyzed using a Perl script. Our main goal was to find the most frequent associations for each word so we calculated the number of times the same association was written for the word. For example, for the word “bucurie” (joy) 18 of 50 respondents indicated “fericire” (happiness), 7 respondents indicated “zâmbet” (smile), 6 respondents indicated “veselie” (fun), other associations were different and had frequency less than 3. Thus the strongest associations for the word “bucurie” (joy) were “fericire” (happiness), “zâmbet” (smile) and “veselie” (fun). We preserved all the associations provided even those with the frequency equal to one keeping in mind the aim to enlarge our associative dictionary.

The overall results are presented in the table 1 and figures 1, 2. Table 1 contains 12 most frequent pairs of stimulating word and associated word. For example, the pair “forța - putere” ("force-power") has the highest frequency: 29 respondents provided this association. In general a great number of associations were synonyms or near-synonyms (9). Even if the association was not synonym as in example “bucătărie - mîncare” (kitchen - food) the association in most cases the same part of speech as the stimulating word. There is a small number of exceptions as, for example, “piatra - tare” (stone - hard).

4. The second step of creating the Romanian word association database

After the first phase of the dictionary creation, we had 150 words-stimulus and 50 responses for each of these words. This information was organised in MySQL database

which we intended to enlarge. In order to obtain more word-associations we created an online interface for our dictionary using PHP[10]. The interface is presented in the figure 3. It can be accessed on http://lilu.fcim.utm.md/asociere/.

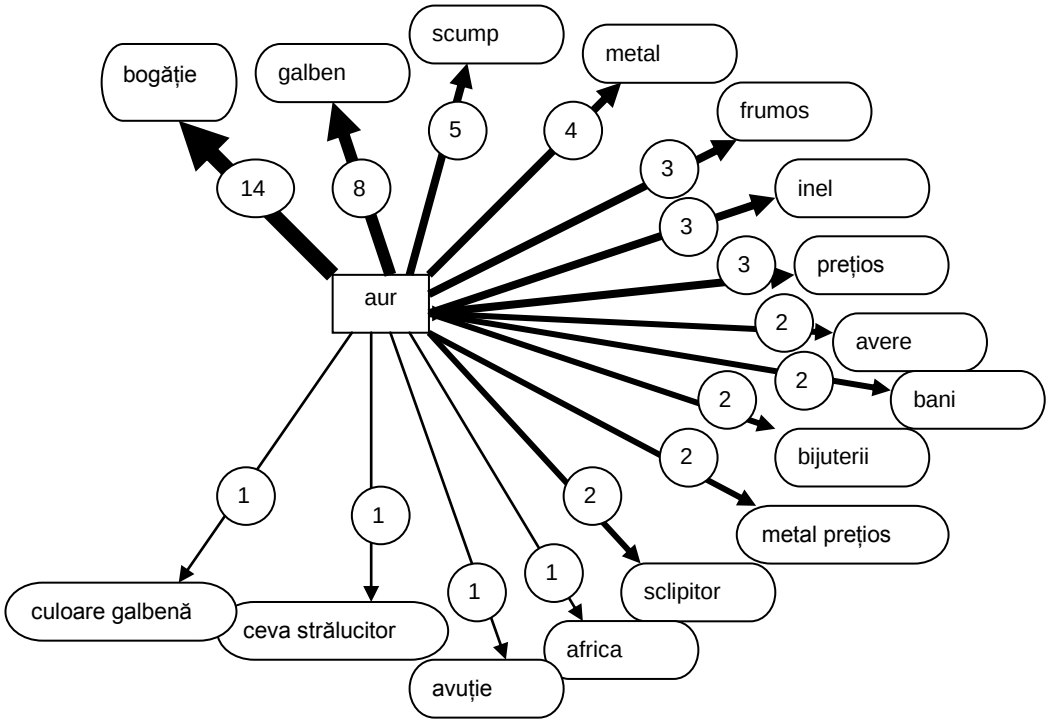

Figure 1: The set associated to the word "gold" that contains 16 associations. The figures on the arrows show the number of respondents who gave this answer.

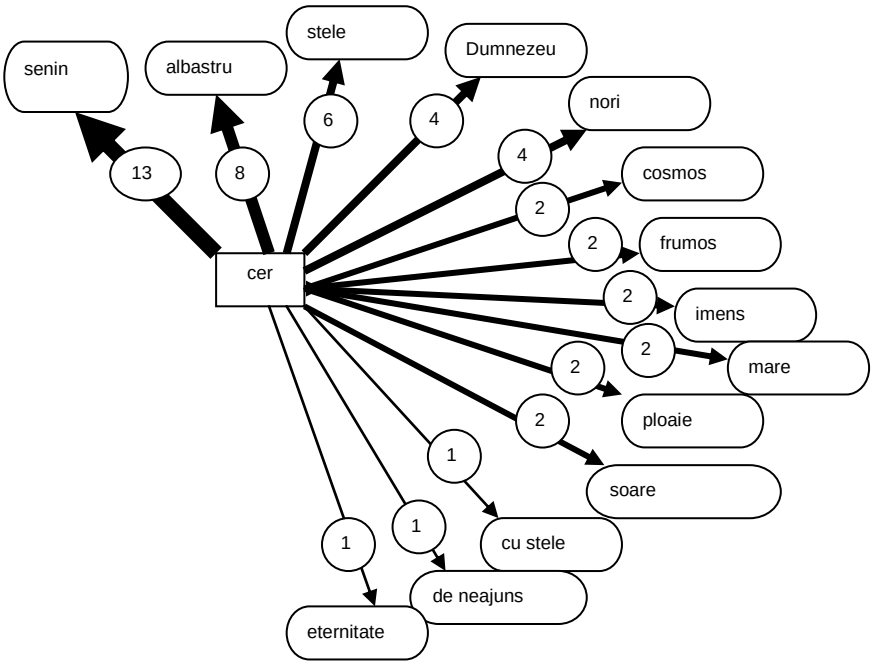

Figure 2: The set associated with the word "sky" which contains 14 associations.

[10] Data base and interface are impemented by our former student Ion Badan.

In order to obtain associations, a user has to type the word in the input field and press the button. An example of result for the word “frumos” (beautiful) is presented in figure 4. The list of all associations is sorted initially by the frequency in descending order; it can also be sorted by any other column of the table, in descending or ascending order.

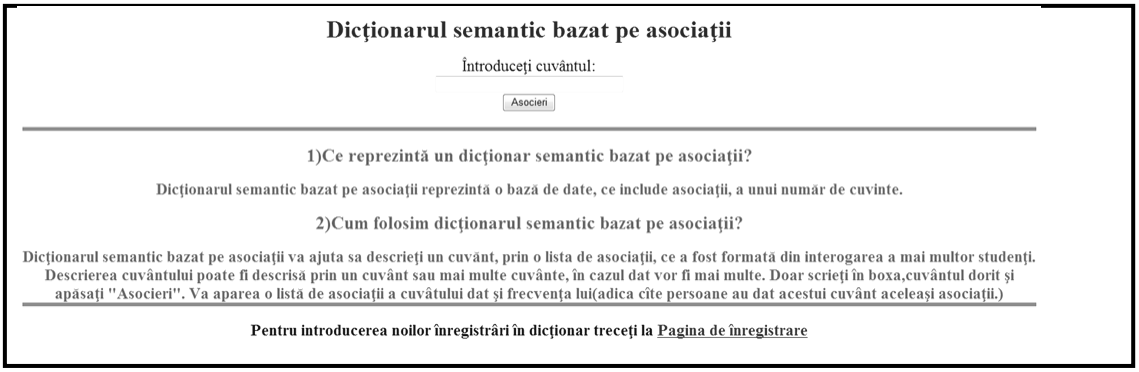

Figure 3: The interface for the Romanian associative dictionary interrogation.

Asocierile pentru cuvântul: “frumos”

#	Asociere ↑/ Asociere ↓	Asociere ↑/ Asociere ↓	Frecvenţa ↑ / Frecvenţa ↓
1	femeie	**frumos**	4
2	**frumos**	placut	4
3	zâmbet	**frumos**	4
4	aur	**frumos**	3
5	corpul	**frumos**	3
6	**frumos**	atragator	3
7	**frumos**	copil	3
8	viitor	**frumos**	3
9	ceas	**frumos**	2
10	cer	**frumos**	2

Rezultate de la 1 la 10 din 66.

1 | 2 | 3 | 4 | 5 | 6 | 7 | Următorul »

Figure 4: The associations for the word “frumos” (beautiful) extracted from Romanian associative dictionary.

There are two types of relations between words in the associative dictionary: *direct relation* from stimulus toward response, and the *inverse relation* from response to stimulus; these relations are not symmetrical. Thus, for the stimulus “aur” (gold) three responses were “frumos” (beautiful), but if “frumos” (beautiful) was stimulus no one response was “aur” (gold).

The resulting table for the interrogation contains both types of relations for the introduced word; it can be seen in figure 4. The first column contains words-stimuli; the second one contains words-responses. The word “frumos” (beautiful) appears in both columns; in the first column as the stimulus and in second as response.

The last line of text in the interface presented in figure 3 contains the link to the page created for introduction of the new records in the associative dictionary. This page is

presented in figure 5. A random word is presented to the user, and the user has to introduce the associated word in the input box. After clicking the button "Asociază", the user is informed that the introduced association was added in the data base. For example:

"Baza de date a fost actualizată cu success pentru înregistrarea lemn <-> foc"

(The database was successfully updated for the record wood<->fire)

Word – stimulus is selected randomly from the list of al words in the database both stimuli and responses. Thus the number of stimuli is also growing more than these 150 words selected initially.

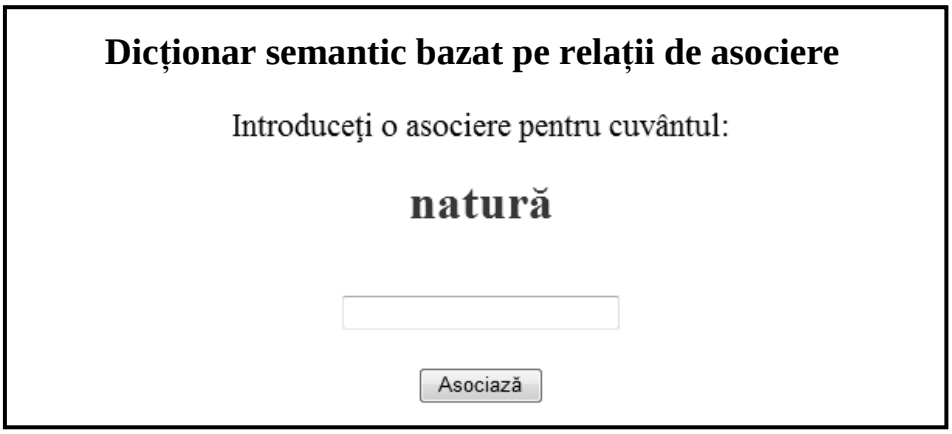

Figure 5: The interface for the Romanian associative dictionary augmentation.

The first version of the association database obtained after processing the questionnaires contained almost 7 500 stimulus-response pairs. We had to remove some of responses for different reasons. Some respondents were not accurate and missed some words, some wrote long phrases instead of words as responses, which we had to remove. After preprocessing we obtained 5821 different pairs; 4152 pairs had frequency equal to 1. Since the database was installed online, it has been augmenting. Statistics until 11 of November 2011 is the following: 10092 pairs total, 6163 different pairs, 4464 pairs had frequency equal to 1.

There are several problems which still remain to be solved. First, the words added online should be verified. A user can add wrong information, a word with grammatical errors or even a combination of letters without any sense. Automatic verification against a dictionary can discard words which are not in our dictionary and if the word is written with a grammatical error, it is extremely difficult to correct it automatically. Diacritic signs represent another problem. Some users introduce words with these signs; some ignore them as it is a usual practice while writing online. The same word typed in two forms, with diacritic signs and without them, is considered as two different words in the database. For example, the stimulus word "zice" (say, speak) has three variants of word "vorbește" (talk) as a response: "vobeste", "vorbeste" and "vorbește". The first one has one letter missed and no diacritics, the second one is correct but without diacritics and the third one is absolutely correct. All of them are stored as three different responses in the current version of the association database.

5. Conclusion

The paper reports about the experiment of an associative dictionary for Romanian language creation. It outlines the first phase of the experiment when *word associations* were collected using questionnaire surveys. The second phase includes online interface creation and expanding of the dictionary via internet. Several technical issues of the second phase are discussed. Some preliminary statistics is presented and a brief analysis of the obtained database is made. The created dictionary can be used in lexicography and Romanian language studying. At this stage of work, we however are more interested in the richest and the most representative database of word association; the detailed analysis is postponed to forthcoming research.

Acknowledgements. The authors are grateful to anonymous reviewers for the profound analysis of our paper and helpful comments.

References

Church, K. W., Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, 22-29.

Collins, A. M., Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407-428.

Cramer, P. (1968). Word association. *New York & London: Academic Press*.

Deese, J. (1965). The structure of associations in language and thought. *Baltimore: The John Hopkins Press*.

Edmonds, P., Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics* 28:2,105-144.

Joyce, T. (2005). Constructing a large-scale database of Japanese word associations. *Special issue edited by Katsuo Tamaoka: Corpus Studies on Japanese Kanji*. Glottometrics, 10, 82-98.

Hirst, G. (2004). Ontology and the lexicon. *In Steffen Staab, & Rudi Studer, (Eds.), Handbook of ontologies*. Berlin, Heidelberg, & New York: Springer-Verlag, 209-229.

Kiss, G. R., Armstrong, C., Milroy, R., Piper, J. (1973). An associative thesaurus of English and its computer analysis. *In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (Eds.), The Computer and Literary Studies. Edinburgh: Edinburgh University Press*.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38: 11, 39-41.

Nelson, D. L., McEvoy, C. L., Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/ FreeAssociation/.

Pustejovsky, J. (2010). Qualia Roles. The Cambridge Encyclopedia of the Language Sciences. *Ed. Patrick Hogan*, Cambridge, UK: Cambrige University Press.

Steyvers, M., Shiffrin, R. M., Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *In A. F. Healy, (Ed.),*

Experimental cognitive psychology and its applications. (Decade of behavior). Washington, D.C.: American Psychological Association, 237-249.

Vlad, A., Mitrea, A., Mitrea, M. (2005). Limba română scrisă ca sursă de informație. *Paideia, România*.

Балтова, П., Ефтимова, А., Липовска, А., Петрова, К. (2003). БАС 2003: Български асоциативен речник. *София: Изд. СУ* "Св. Кл. Охридски".

Караулов, Ю. Н., Черкасова, Г. А., Уфимцева, Н. В., Сорокин, Ю. А., Ярошинская, В. Н. (2002, 2003). РУССКИЙ АССОЦИАТИВНЫЙ СЛОВАРЬ. Том I. От стимула к реакции. Том II. От реакции к стимулу. Астрель, АСТ, 784 (992) стр.

Уфимцева, Н. В. (2004). Славянский ассоциативный словарь: русский, белорусский, болгарский, украинский. *Институт языкознания РАН*, 790 стр.